W458–W465 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 8 May 2012*
doi:10.1093/nar/gks380

GPSy: a cross-species gene prioritization system for conserved biological processes—application in male gamete development

Ramona Britto[1], Olivier Sallou[2], Olivier Collin[2], Grégoire Michaux[3], Michael Primig[1] and Frédéric Chalmel[1,*]

[1]Inserm Unité 1085-Irset, [2]Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA/INRIA) - GenOuest platform and [3]CNRS, UMR6061, Université de Rennes 1, F-35043 Rennes, France

Received January 27, 2012; Revised March 30, 2012; Accepted April 12, 2012

ABSTRACT

We present gene prioritization system (GPSy), a cross-species gene prioritization system that facilitates the arduous but critical task of prioritizing genes for follow-up functional analyses. GPSy's modular design with regard to species, data sets and scoring strategies enables users to formulate queries in a highly flexible manner. Currently, the system encompasses 20 topics related to conserved biological processes including male gamete development discussed in this article. The web server-based tool is freely available at http://gpsy.genouest.org.

INTRODUCTION

High-throughput technologies have generated a vast amount of biological information. However, it remains a difficult task for biologists and clinical researchers to identify genes potentially important for a given biological process or disorders related to it based on these data. When various sources of information are weighted and prioritized by investigators based on their subjective perception of how important they are, a bias may be introduced. To tackle this critical problem, the bioinformatics field has developed a number of solutions for gene prioritization (1); these methods are typically based on the idea that genes whose expression patterns, subcellular localization, structural domains, molecular functions or physical interactions are similar to those known to be important for a given biological process or a pathology, are likely to play critical roles as well. Alternatively, genes can be prioritized on the basis of domain-specific knowledge for specific diseases and biological processes (2,3). The tools available are either standalone applications (4–6) or solutions implemented on web servers (1). These systems exploit several data sources and many of them require known ('training') genes as a control (positive) reference set for prioritization (1,7–12). A number of these solutions bring together information from diverse sources both within and across species and are often too vast to be integrated manually. The existing solutions, while very useful, are limited in the choice of species, query options and coverage of data types. Moreover, none of them fully exploit multiple sources of information across species.

The majority of existing approaches (Supplementary Table S1) are centered on human, some include several species (13–16), and others utilize data from one organism to drive prioritization in another species (4,11–13,17–22). Chen *et al.* (11) demonstrated that the inclusion of a single data type (phenotype) from an alternate organism (mouse) significantly improved prioritization of human disease candidates. Protein–protein interaction data from multiple organisms has also been shown to aid gene prioritization (12,21,22). This cross-species capability, however, is restricted to a single data type in each case.

Our lab has been developing and maintaining solutions for genome biological data management, data analysis and data dissemination (23–25) during the last decade. Here, we present the first release of the gene prioritization system (GPSy), which currently covers 20 topics related to conserved biological processes including cellular development and differentiation (3 topics), organ/tissue development (15 topics) and disorders/diseases (2 topics; Supplementary Table S2 for a complete list). Users can query the system with genes from a list of 45 eukaryotic species including all major model organisms; it is possible to upload lists of genes identified via expression profiling, proteomics, genome wide association (GWA) studies or even complete genomes. The submitted lists of genes are analysed using biological data falling into four broad categories (Sequence, Expression, Annotation and Association) each in combination with a specific ranking method (Figure 1A and Supplementary Table S3).

*To whom correspondence should be addressed. Tel: +33 2 23 23 58 02; Fax: +33 2 23 23 50 55; Email: frederic.chalmel@inserm.fr

Importantly, the ranking parameters are flexible which enables users to attribute different weights and to select species of interest for each data type (Figure 1B). We provide an optimized weight scheme for each topic based on an evaluation of different weight combinations ranging from 1 to 10 for each data type. Taken together, these features allow for complex queries pertaining to very specific questions for each topic. We have successfully tested GPSy using worm homologs of mammalian candidate genes followed by validation using phenotypic data from high-throughput RNA interference (RNAi) studies in *Caenorhabditis elegans* (26) and our own manual RNAi experiments.

GPSy is thus the first system that integrates a large variety of data across a wide range of organisms. GPSy's approach to gene prioritization makes it a tool that is applicable to many different fields, in particular, those focussing on conserved biological processes and their related disorders.

RESULTS

User interface: data input/data output

GPSy has a simple and intuitive interface including a Query tab which enables users to first select one of 20 topics that are currently available from a dropdown menu and then to define the query species. A text field is available to enter the list of candidates; alternatively, the user can request prioritization of 1000 random selected genes or the entire genome for the chosen species. Additionally, for human, a set of positive reference genes can be uploaded for each topic. Currently, GPSy only accepts Entrez Gene identifiers (IDs) because reliable and consistent gene ID conversion is a complex problem; users are referred to two up-to-date resources for gene ID unification over a wide range or organisms (27,28). It is possible to select individual species and data modules and to modify their weights (from 0 to 10) using the Advanced options tab (Figure 1B). By default, all data sets are selected for all available species ($n = 45$) and the preset parameters from the optimal weight scheme are applied.

The output page displays the top 50 genes by default but users can change this setting as they deem appropriate. The result is displayed in the form of a table containing one gene per line with columns for Gene IDs (hyperlinked to the NCBI), Priority ranking, individual module ranks and other relevant information. The weight used in each module to compute the overall score is indicated in brackets. The output list is ordered (prioritized) according to the overall score; it can be reordered based on the ranks of individual modules. Information regarding the intra-module ranks is accessible through the magnifying glass icon. The table in the html output displays the top 1000 genes; the entire gene list and corresponding ranking information can be exported as an archive file (.tar) via the 'Export results' link at the bottom of the page. The welcome page includes a link to a brief tutorial for GPSy.

Species and homology

We assembled a map of conserved genes across the 45 eukaryotic species for which complete genome sequence information was available (Supplementary Table S3). Related homolog clusters from NCBI's HomoloGene (29) and the OMA (Orthologous MAtrix) (30) projects were merged using verified homolog pairs (BLAST reciprocal best hits) as suggested by Roth *et al.* (31) Supplementary Figure S2A).

Modules and ranking

Thirteen different types of genomic data common to the included topics were assembled from various sources (Supplementary Table S1). These were organized into four data categories: Sequence, Expression, Annotation and Association each associated with a unique scoring strategy. The integration of genome data sets with distinct scoring strategies forms the basis of GPSy's modular architecture allowing for maximum query flexibility (Figure 1A). The choice of data sources and scoring strategies is explained in detail in Supplementary Methods. In contrast to methods used in generic gene prioritization tools, the process-specific approach implemented in GPSy enables the pre-computation of module- and species-wise ranks; a feature that greatly accelerates the process of prioritization.

When the system is queried, candidate genes in the input list are mapped onto the pre-computed ranked lists for the corresponding species. An intra-module weighted average rank is computed for each gene in the input list by combining the relative ranks for the input species according to every other selected species.

Positive and negative reference gene sets

Positive reference sets (PRSs) of genes known to be relevant for each topic were assembled for the 45 species and used for scoring genes in the Annotation and Association categories (Supplementary Table S5). For this purpose, information was gathered from the Gene Ontology and phenotype projects in various organisms. The ontological structure of these data allowed us to identify the ensemble of relevant annotation terms for each topic. This included 'biological process' terms from the Gene Ontology project (e.g. gamete generation) and species-specific phenotype terms (e.g. azoospermia; listed in Supplementary Table S4). Negative reference sets (NRSs) of 1000 randomly chosen genes not annotated with the selected terms were generated as controls. Note that the human PRS and NRS were employed in the Weightage optimization procedure.

Weightage optimization and overall prioritization

To assess the contributions of each module to overall prioritization, we decided to test the effectiveness of different weight combinations. We employed an approach similar to Sun *et al.* (2), to test different weight vectors (ranging from 1 to 10) in the 13 different modules for each topic (Supplementary Table S2). To evaluate the performance

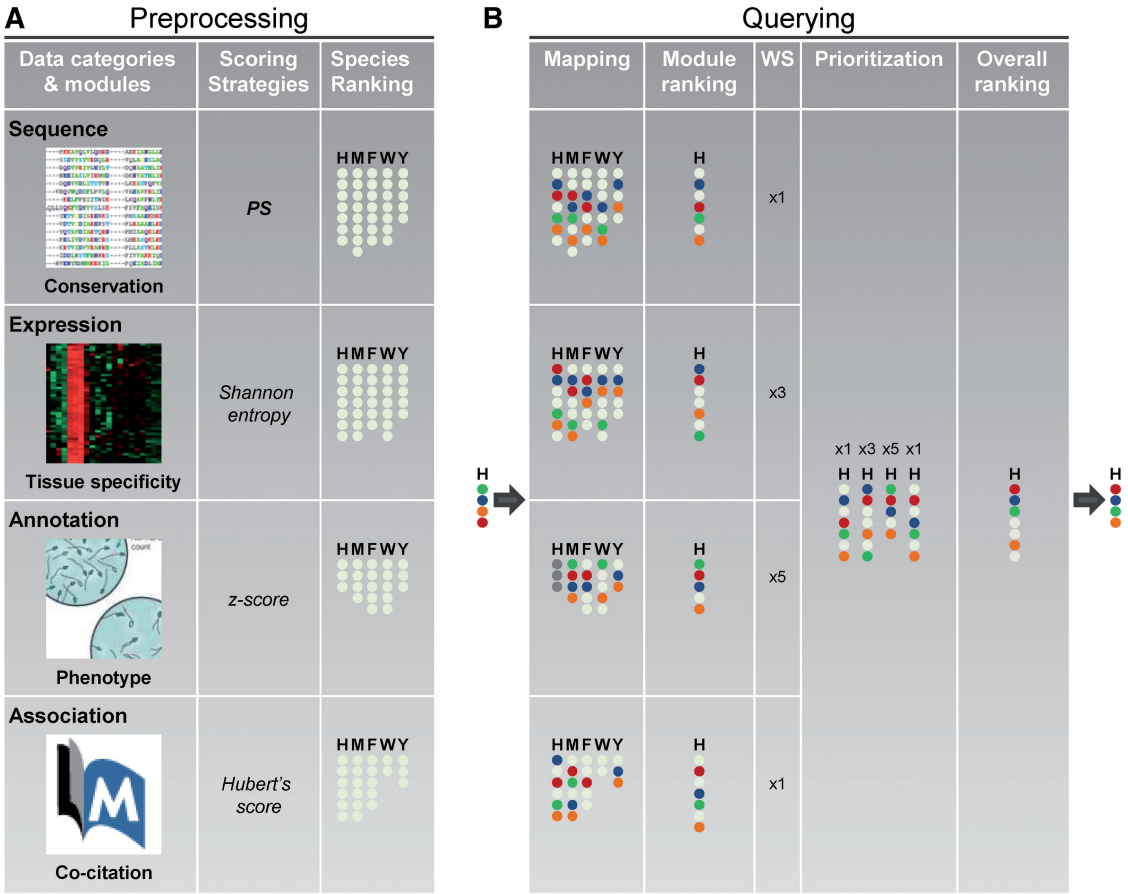

Figure 1. Framework for the prioritization of candidate genes. (**A**) and (**B**) describe the steps involved in pre-processing and querying respectively. Lane 1 (Data categories and modules) lists a non-exhaustive list of modules falling into the four categories (Sequence, Expression, Annotation and Association) that were collected and curated from different species to drive gene prioritization. Lane 2 outlines the scoring strategies, one for each module. The species-wise ranking process that follows the scoring of individual genes is depicted in Lane 3. H, M, F, W and Y indicate the ranked lists for human, mouse, fly, worm and yeast, respectively. (B) The server accepts as input a gene list from any one of the 45 species (human, in the displayed example). Genes in the input list are mapped onto pre-computed ranked lists for selected species (Lane 4) and an intra-module rank is generated (Lane 5). Lane 6 (WS; Weight Scheme) highlights the weight applied to each module. Lanes 7 and 8 describe the final step in gene prioritization, calculation of an inter-module weighted average rank for each gene. The output is the prioritized input list.

of each weight combination, a discrimination analysis method was employed. Sensitivity and specificity values were computed and a receiver operating characteristic (ROC) curve was plotted (1-Specificity versus Sensitivity). The area under this curve (AUC) corresponds to the probability that a random positive instance will score higher than a random negative instance (32). An AUC of 1 indicates that all PRS genes ranked above NRS genes; 0.5 indicates that the genes ranked randomly. As an exhaustive test of all weight combinations (2) is impractical (10^{13} weight schemes), we employed a heuristic approach to achieve a satisfactory discrimination of true positives (PRS) from true negative (NRS) candidates (Supplementary Methods). The overall rank of a given gene is an inter-module weighted average of the individual module ranks. The final output is a reordered list based on the overall ranking of each gene. A more detailed description of the pre-processing steps and overall prioritization can be found in Supplementary Methods.

Caenorhabditis elegans as a model for spermatogenesis

The worm is a key model organism for the high-throughput analysis of genes involved in meiotic development; these functional studies typically involve small interfering RNA (siRNA) which down-regulates mRNA expression (33). High-throughput RNAi studies are informative; however, they are often limited to detecting specific defects and are biased by a number of experimental artefacts such as wrongly annotated RNAi clones and false-positive or false-negative phenotype scores. Finally, the penetrance of a phenotype depends upon the technique used: RNAi feeding where worms are bred on a layer of bacteria containing a plasmid expressing the siRNA is less efficient than direct RNAi injection or the use of a bona fide gene deletion strain. To corroborate GPSy's ranking output, we therefore decided to test the ability of a selected group of genes to induce a sterility or germ line defect phenotype in a strain background particularly sensitive to RNAi by the feeding method (Supplementary File S5).

We first selected 56 *C. elegans* orthologues of mammalian genes previously identified in our lab as strongly induced in the worm and mouse germ line (34). Among the 56 genes investigated, 23 were associated with a reproductive phenotype (RP corresponding to sterility or a germ line defect) when the union of results from our RNAi experiments (11 genes associated with RP; Supplementary File S4) and those of large-scale and individual studies available via Wormbase (18 genes associated with RP) were taken into consideration. These additional phenotypes reported but not identified in our experiments are likely due to different strain backgrounds and experimental approaches. The remaining 33 genes (non-RP set) showed no clearly detectable RP under the conditions we and others employed. Next, we prioritized the worm gene list (56 genes) using GPSy's Spermatogenesis topic using default weight settings and all species and modules with the exception of *C. elegans* phenotype data. The output list was integrated with phenotypic information from our and other experiments (23 RP and 33 non-RP genes; Figure 2A).

Combining the GPSy ranks with the validated phenotypic data suggests a promising pattern, we observe a tendency for genes associated with reproductive phenotypes (RP phenotype class) to receive a high rank in comparison to genes whose involvement in the gametogenic process could not be established (bottom of the list, non-RP classes; Figure 2A). Eight of the top 10 genes display a reproductive or lethal phenotype. These genes are discussed in Supplementary File S5. The lower half of the list has relatively few genes with documented germ line/sterility phenotypes. The overall trend for high-ranking genes to result in a sterility/germ line defect phenotype is also demonstrated by the reliable discrimination of genes associated a reproductive phenotype (RP, $n = 23$) from a worm negative reference set (NRS, $n = 1000$) based on GPSy ranking (Figure 2B). Since the candidate list ($n = 56$) itself is expected to be enriched for PRS genes, its AUC is non-random (75.2%). This is, however, significantly lower than the AUC obtained with RP genes alone (86.2%). The ranking also demonstrated sufficient discriminability within the candidate list (RP versus non-RP; AUC = 71.9%). A chi-square test performed on the same set (RP genes against all others) revealed a statistically significant trend ($P = 0.002$).

To illustrate the contribution of cross-species information, we subjected the gene list to GPSy prioritization without considering data from homologs in other species. The resulting difference in AUC value (0.582 versus 0.722) clearly illustrates the value of the cross-species approach (Figure 2C).

Comparison to other methods

We wanted to test GPSy's ability to efficiently prioritize the worm candidate gene list in comparison to existing approaches. A comprehensive survey of freely available, web-based gene prioritization software revealed that for *C. elegans*, as with most non-human species, the choices are limited (Supplementary Table S1). Seven of the 30 tools compared offer multi-species capability. Of these, only two tools allow the querying of *C. elegans* data sets and provide gene ranking based on diverse data types thus enabling comparison with GPSy's results. The performance of these two tools, Génie and Endeavour (13,16), was compared to that of GPSy using the discrimination analysis method described. We subjected the *C. elegans* shortlist ($n = 56$) to GPSy and to Endeavour using default parameters. We used the worm PRS for spermatogenesis as the training set for Endeavour. For Génie, we used 'spermatogenesis' as topic of interest, a *P*-value cutoff of 1.0 for abstracts and a false discovery rate of 1.0 for gene selection, while taking into consideration all possible orthologues. The resulting receiver operating characteristic (ROC) curves and corresponding AUC values show significant differences among the tools in favor of GPSy (72.2%) as compared to Génie (68.9%) and Endeavour (65.2%; Figure 2C). We also observed a considerable increase in computation time for the method dependent on a training set (~10 min using Endeavour as against 10 s for GPSy). The justification of several high- and low-ranking genes obtained through a fair validation strategy (exclusion of worm phenotype data during prioritization), point to the effectiveness of the cross-species approach. The correlation of GPSy rank and phenotype relevance (Figure 2A) and the reliable discrimination of genes with and without the phenotype of interest (Figure 2B and C), suggest that the use of this system on large candidate gene lists will enable the focusing of time and experimental resources on those predictions most likely to be true.

DISCUSSION

The wide variety of data types included in GPSy, in conjunction with its modular nature, enables users to address very specific biological questions. In the Spermatogenesis topic, maximizing the weight of the Tissue specificity module may be advantageous for identifying potential gonad (germ line)-specific marker genes across species. On the other hand, decreasing the weight of Gene Ontology and Phenotype annotations for the query species, improves the ranking of uncharacterized genes, thus facilitating the discovery of novel genes important for the selected topic.

In comparison to other prioritization methods, GPSy covers many more data sources and provides users with a choice of different species (Supplementary Table S1). The multi-species capability is important for basic scientists whose research is primarily conducted in model organisms. This feature is especially valuable for recently sequenced organisms and others where little or no data beyond the genomic sequence are available (27 out of 45 species; Supplementary Table S3). The value of a cross-species approach is evident also in the case of established model organisms; for example, very little phenotype/disease data are available for primates in comparison to mouse, fly, worm and yeast.

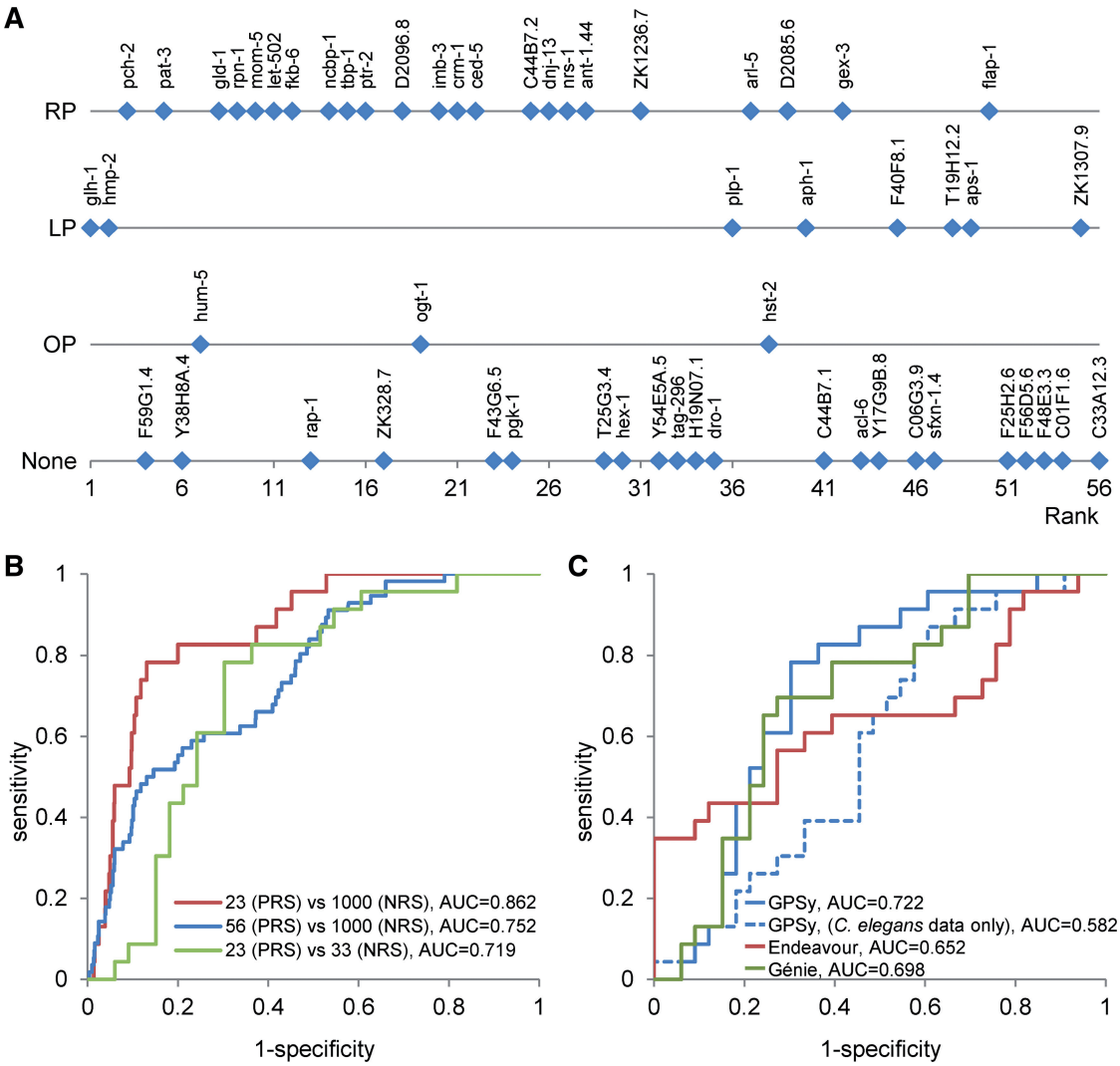

Figure 2. Gene ranking and RNAi phenotypes. (**A**) The most relevant phenotypes are plotted for each gene in the prioritized candidate list (from the 1st to the 56th, *x*-axis). On the *y*-axis, phenotype classes are indicated: RP = reproduction-associated phenotype; LP = lethal phenotype; OP = other phenotype; None = no observable phenotype. Official gene symbols are displayed for all genes. (**B**) Displays receiver operating characteristic (ROC) curves for: (i) the candidate gene set ($n = 56$ genes) versus the *C. elegans* negative reference set (NRS; $n = 1000$; blue curve); (ii) the RP genes set ($n = 23$) versus NRS (red); (iii) the RP versus non-RP sets (union of LP, OP and None phenotype; $n = 33$; green). The corresponding area under the ROC curve (AUC) values are indicated. Note the significant improvement in AUC value between (ii) and (i). The AUC value for (iii) is significantly non-random. (**C**) Displays ROC curves for the discrimination of the *C. elegans* RP ($n = 23$) versus non-RP sets ($n = 33$) using GPSy (default settings, solid blue line), GPSy (*C. elegans* data only, dashed blue line), Endeavour (red) and Génie (green).

Existing approaches using machine learning (35), and kernel- (16) or network-based (32,36) strategies generally rely on training gene sets provided during the query. Systems such as GPSy that use pre-defined criteria and pre-computed scores have the advantage of being much faster. GPSy returns priority lists for the mouse and human genomes in 45 s in comparison to 30 min on average in the case of Endeavour (with a small training set and all data sets selected). With the majority of tools, limitations exist for the size of the reference or candidate gene sets, or both; thus a direct comparison of all performance aspects is not feasible.

The choice of positive reference genes (PRS) for training purposes is a critical factor because both the size and the homogeneity of the reference set affect the reliability of gene prioritization. There is usually an inverse relationship between them; for very small training sets, homogeneity increases but at the cost of statistical validity. It has been noted that the training set homogeneity is an important factor for effective ranking (10). Estimating homogeneity is a non-trivial task and the time required for the process increases with the size of the reference set. GPSy uses a comprehensive reference set (PRS) relevant for each topic that was carefully selected and then reviewed by experts in the field. Nevertheless, such contrasting features between GPSy and the other gene prioritization approaches suggest that the tools may be used in a complementary fashion (37).

The effective prioritization of *C. elegans* genes through data available in other species shows that the system is scientifically sound and stresses the importance of a cross-species approach. It is obvious, however, that investigator discretion is important in the inclusion/exclusion of selected species particularly for widely divergent clades (e.g. Human–Plant).

CONCLUSION

We report the development and application of GPSy, a novel multi-dimensional tool which integrates distinct data types across a wide range of organisms. This tool is intended for the rapid identification of genes potentially important for conserved biological processes such as male gamete development. GPSy is modular and extendable which enables us and others to include novel topics and data sets as the need arises. In the future, GPSy will include less utilized datasets such as regulation by non-coding RNAs (38) and others, as they become available. A future release of our tool will include an update of GPSy's 'Cancer' topic through the inclusion of gene expression data in normal versus cancer samples. We intend to complete GPSy's repertoire with other topics of interest related to conserved biological processes in the near future.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–7, Supplementary Figures 1–5, Supplementary Methods, Supplementary Files 1–5 and Supplementary References [39–71].

ACKNOWLEDGEMENTS

We thank members of the laboratory and Inserm Unit 1085 for stimulating discussions, and the GenOuest platform for hosting the software.

FUNDING

Funding for open access charge: Inserm, Région Bretagne (PhD fellowship); University of Rennes 1 awarded (to R.B.); Inserm Avenir [R07216NS to M.P.].

Conflict of interest statement. None declared.

REFERENCES

1. Tranchevent,L.C., Capdevila,F.B., Nitsch,D., De Moor,B., De Causmaecker,P. and Moreau,Y. (2011) A guide to web tools to prioritize candidate genes. *Brief. Bioinformatics*, **12**, 22–32.
2. Sun,J., Jia,P., Fanous,A.H., Webb,B.T., van den Oord,E.J., Chen,X., Bukszar,J., Kendler,K.S. and Zhao,Z. (2009) A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases-schizophrenia as a case. *Bioinformatics*, **25**, 2595–6602.
3. Gajendran,V.K., Lin,J.R. and Fyhrie,D.P. (2007) An application of bioinformatics and text mining to the discovery of novel genes related to bone biology. *Bone*, **40**, 1378–1388.
4. Gaulton,K.J., Mohlke,K.L. and Vision,T.J. (2007) A computational system to select candidate genes for complex human traits. *Bioinformatics*, **23**, 1132–1140.
5. Ma,X., Lee,H., Wang,L. and Sun,F. (2007) CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics*, **23**, 215–221.
6. Morrison,J.L., Breitling,R., Higham,D.J. and Gilbert,D.R. (2005) GeneRank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics*, **6**, 233.
7. Hristovski,D., Peterlin,B., Mitchell,J.A. and Humphrey,S.M. (2005) Using literature-based discovery to identify disease candidate genes. *Int. J. Med. Inform.*, **74**, 289–298.
8. Van Vooren,S., Thienpont,B., Menten,B., Speleman,F., De Moor,B., Vermeesch,J. and Moreau,Y. (2007) Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations. *Nucleic Acids Res.*, **35**, 2533–2543.
9. Yu,W., Wulf,A., Liu,T., Khoury,M.J. and Gwinn,M. (2008) Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases. *BMC Bioinformatics*, **9**, 528.
10. Aerts,S., Lambrechts,D., Maity,S., Van Loo,P., Coessens,B., De Smet,F., Tranchevent,L.C., De Moor,B., Marynen,P., Hassan,B. *et al.* (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol.*, **24**, 537–544.
11. Chen,J., Xu,H., Aronow,B.J. and Jegga,A.G. (2007) Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics*, **8**, 392.
12. Kohler,S., Bauer,S., Horn,D. and Robinson,P.N. (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.
13. Fontaine,J.F., Priller,F., Barbosa-Silva,A. and Andrade-Navarro,M.A. (2011) Genie: literature-based gene prioritization at multi genomic scale. *Nucleic Acids Res.*, **39**, W455–W461.
14. Xiong,Q., Qiu,Y. and Gu,W. (2008) PGMapper: a web-based tool linking phenotype to genes. *Bioinformatics*, **24**, 1011–1013.
15. Warde-Farley,D., Donaldson,S.L., Comes,O., Zuberi,K., Badrawi,R., Chao,P., Franz,M., Grouios,C., Kazi,F., Lopes,C.T. *et al.* (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, **38**, W214–W220.
16. Tranchevent,L.C., Barriot,R., Yu,S., Van Vooren,S., Van Loo,P., Coessens,B., De Moor,B., Aerts,S. and Moreau,Y. (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res.*, **36**, W377–W384.
17. Yoshida,Y., Makita,Y., Heida,N., Asano,S., Matsushima,A., Ishii,M., Mochizuki,Y., Masuya,H., Wakana,S., Kobayashi,N. *et al.* (2009) PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning. *Nucleic Acids Res.*, **37**, W147–W152.
18. Seelow,D., Schwarz,J.M. and Schuelke,M. (2008) GeneDistiller: distilling candidate genes from linkage intervals. *PLoS One*, **3**, e3874.
19. Yue,P., Melamud,E. and Moult,J. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166.
20. Hutz,J.E., Kraja,A.T., McLeod,H.L. and Province,M.A. (2008) CANDID: a flexible method for prioritizing candidate genes for complex human traits. *Genet. Epidemiol.*, **32**, 779–790.
21. George,R.A., Liu,J.Y., Feng,L.L., Bryson-Richardson,R.J., Fatkin,D. and Wouters,M.A. (2006) Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res.*, **34**, e130.
22. Franke,L., van Bakel,H., Fokkens,L., de Jong,E.D., Egmont-Petersen,M. and Wijmenga,C. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, **78**, 1011–1025.
23. Chalmel,F. and Primig,M. (2008) The annotation, mapping, expression and network (AMEN) suite of tools for molecular systems biology. *BMC Bioinformatics*, **9**, 86.
24. Gattiker,A., Hermida,L., Liechti,R., Xenarios,I., Collin,O., Rougemont,J. and Primig,M. (2009) MIMAS 3.0 is a multiomics

information management and annotation system. *BMC Bioinformatics*, **10**, 151.
25. Lardenois,A., Chalmel,F., Barrionuevo,F., Demougin,P., Scherer,G. and Primig,M. (2010) Profiling spermatogenic failure in adult testes bearing Sox9-deficient Sertoli cells identifies genes involved in feminization, inflammation and stress. *Reprod. Biol. Endocrinol.*, **8**, 154.
26. Harris,T.W., Antoshechkin,I., Bieri,T., Blasiar,D., Chan,J., Chen,W.J., De La Cruz,N., Davis,P., Duesbury,M., Fang,R. *et al.* (2010) WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.*, **38**, D463–D467.
27. Baron,D., Bihouee,A., Teusan,R., Dubois,E., Savagner,F., Steenman,M., Houlgatte,R. and Ramstein,G. (2011) MADGene: retrieval and processing of gene identifier lists for the analysis of heterogeneous microarray datasets. *Bioinformatics*, **27**, 725–726.
28. Chen,R., Li,L. and Butte,A.J. (2007) AILUN: reannotating gene expression data automatically. *Nat. Methods*, **4**, 879.
29. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Federhen,S. *et al.* (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **40**, D13–D25.
30. Altenhoff,A.M., Schneider,A., Gonnet,G.H. and Dessimoz,C. (2011) OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res.*, **39**, D289–D294.
31. Roth,A.C., Gonnet,G.H. and Dessimoz,C. (2008) Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics*, **9**, 518.
32. Tom,F. (2006) An introduction to ROC analysis. *Pattern Recogn. Lett.*, **27**, 861–874.
33. Timmons,L. and Fire,A. (1998) Specific interference by ingested dsRNA. *Nature*, **395**, 854.
34. Chalmel,F., Rolland,A.D., Niederhauser-Wiederkehr,C., Chung,S.S., Demougin,P., Gattiker,A., Moore,J., Patard,J.J., Wolgemuth,D.J., Jegou,B. *et al.* (2007) The conserved transcriptome in human and rodent male gametogenesis. *Proc. Natl Acad. Sci. USA*, **104**, 8346–8351.
35. Adie,E.A., Adams,R.R., Evans,K.L., Porteous,D.J. and Pickard,B.S. (2005) Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, **6**, 55.
36. Xu,J. and Li,Y. (2006) Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics*, **22**, 2800–2805.
37. Thornblad,T.A., Elliott,K.S., Jowett,J. and Visscher,P.M. (2007) Prioritization of positional candidate genes using multiple web-based software tools. *Twin Res. Hum. Genet.*, **10**, 861–870.
38. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
39. Kuzniar,A., van Ham,R.C., Pongor,S. and Leunissen,J.A. (2008) The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.*, **24**, 539–551.
40. Altenhoff,A.M. and Dessimoz,C. (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput. Biol.*, **5**, e1000262.
41. Baudat,F., Manova,K., Yuen,J.P., Jasin,M. and Keeney,S. (2000) Chromosome synapsis defects and sexually dimorphic meiotic progression in mice lacking Spo11. *Mol. Cell*, **6**, 989–998.
42. Klapholz,S., Waddell,C.S. and Esposito,R.E. (1985) The role of the SPO11 gene in meiotic recombination in yeast. *Genetics*, **110**, 187–216.
43. Romanienko,P.J. and Camerini-Otero,R.D. (2000) The mouse Spo11 gene is required for meiotic chromosome synapsis. *Mol. Cell*, **6**, 975–987.
44. Muller,J., Creevey,C.J., Thompson,J.D., Arendt,D. and Bork,P. (2010) AQUA: automated quality improvement for multiple sequence alignments. *Bioinformatics*, **26**, 263–265.
45. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.
46. UniProt Consortium. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.
47. Turner,F.S., Clutterbuck,D.R. and Semple,C.A. (2003) POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol.*, **4**, R75.
48. Nitsch,D., Tranchevent,L.C., Goncalves,J.P., Vogt,J.K., Madeira,S.C. and Moreau,Y. (2011) PINTA: a web server for network-based gene prioritization from expression data. *Nucleic Acids Res.*, **39**, W334–W338.
49. Masotti,D., Nardini,C., Rossi,S., Bonora,E., Romeo,G., Volinia,S. and Benini,L. (2008) TOM: enhancement and extension of a tool suite for in silico approaches to multigenic hereditary disorders. *Bioinformatics*, **24**, 428–429.
50. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets: 10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.
51. Parkinson,H., Sarkans,U., Kolesnikov,N., Abeygunawardena,N., Burdett,T., Dylag,M., Emam,I., Farne,A., Hastings,E., Holloway,E. *et al.* (2011) ArrayExpress update: an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **39**, D1002–D1004.
52. Primig,M., Williams,R.M., Winzeler,E.A., Tevzadze,G.G., Conway,A.R., Hwang,S.Y., Davis,R.W. and Esposito,R.E. (2000) The core meiotic transcriptome in budding yeasts. *Nat. Genet.*, **26**, 415–423.
53. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
54. Reinke,V., Gil,I.S., Ward,S. and Kazmer,K. (2004) Genome-wide germline-enriched and sex-biased expression profiles in Caenorhabditis elegans. *Development*, **131**, 311–323.
55. Schug,J., Schuller,W.P., Kappen,C., Salbaum,J.M., Bucan,M. and Stoeckert,C.J. Jr (2005) Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.*, **6**, R33.
56. Rogers,M.F. and Ben-Hur,A. (2009) The use of gene ontology evidence codes in preventing classifier assessment bias. *Bioinformatics*, **25**, 1173–1177.
57. Gene Ontology Consortium. (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.*, **38**, D331–D335.
58. Matzuk,M.M. and Lamb,D.J. (2008) The biology of infertility: research advances and clinical challenges. *Nat. Med.*, **14**, 1197–1213.
59. Davis,A.P., King,B.L., Mockus,S., Murphy,C.G., Saraceni-Richards,C., Rosenstein,M., Wiegers,T. and Mattingly,C.J. (2011) The Comparative Toxicogenomics Database: update 2011. *Nucleic Acids Res.*, **39**, D1067–D1072.
60. Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.
61. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.
62. Doniger,S.W., Salomonis,N., Dahlquist,K.D., Vranizan,K., Lawlor,S.C. and Conklin,B.R. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.*, **4**, R7.
63. Gentleman,R., Scholtens,D., Ding,B., Carey,V.J. and Huber,W. (2005) Graph Case Studies: Literature co-citation. In: Gentleman,R., Carey,V.J., Huber,W., Irizarry,R.A. and Dudoit,S. (eds), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer-Verlag, pp. 378–387, http://www.bioconductor.org/help/publications/tech-reports/.
64. Saccone,S.F., Saccone,N.L., Swan,G.E., Madden,P.A., Goate,A.M., Rice,J.P. and Bierut,L.J. (2008) Systematic biological prioritization after a genome-wide association study: an application to nicotine dependence. *Bioinformatics*, **24**, 1805–1811.

65. Liekens,A.M., De Knijf,J., Daelemans,W., Goethals,B., De Rijk,P. and Del-Favero,J. (2011) BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biol.*, **12**, R57.
66. Kamath,R.S. and Ahringer,J. (2003) Genome-wide RNAi screening in Caenorhabditis elegans. *Methods*, **30**, 313–321.
67. Kirino,Y., Vourekas,A., Kim,N., de Lima Alves,F., Rappsilber,J., Klein,P.S., Jongens,T.A. and Mourelatos,Z. (2010) Arginine methylation of vasa protein is conserved across phyla. *J. Biol. Chem.*, **285**, 8148–8154.
68. Hao,Z., Jha,K.N., Kim,Y.H., Vemuganti,S., Westbrook,V.A., Chertihin,O., Markgraf,K., Flickinger,C.J., Coppola,M., Herr,J.C. *et al.* (2004) Expression analysis of the human testis-specific serine/threonine kinase (TSSK) homologues. A TSSK member is present in the equatorial segment of human sperm. *Mol. Hum. Reprod.*, **10**, 433–444.
69. Xu,B., Hao,Z., Jha,K.N., Zhang,Z., Urekar,C., Digilio,L., Pulido,S., Strauss,J.F. III, Flickinger,C.J. and Herr,J.C. (2008) Targeted deletion of Tssk1 and 2 causes male infertility due to haploinsufficiency. *Dev. Biol.*, **319**, 211–222.
70. Korswagen,H.C., Herman,M.A. and Clevers,H.C. (2000) Distinct beta-catenins mediate adhesion and signalling functions in C. elegans. *Nature*, **406**, 527–532.
71. Wu,M. and Herman,M.A. (2006) A novel noncanonical Wnt pathway is involved in the regulation of the asymmetric B cell division in C. elegans. *Dev. Biol.*, **293**, 316–329.